Evaluation of Single-Channel Reverberation Suppression as Preprocessing for Acoustic Event Detection

Christine Baldzer, Benjamin Cauchi, Jens Schröder , Danilo Hollosi

Fraunhofer Institute for Digital Media Technology (IDMT)
Project Group Hearing, Speech and Audio Technology
26129 Oldenburg, Germany
Email: christine.baldzer@idmt.fraunhofer.de

Abstract

Acoustic event detection (AED) is increasingly present in applications such as health monitoring, security or home automation. This paper investigates the influence of reverberation on the performance of an AED system and the potential benefit of applying single-channel spectral reverberation suppression (SRS) as preprocessing for such a system. Experiments, conducted using a dataset of anechoic acoustic events and publicly available datasets of room impulse responses (RIRs) show that the application of SRS as preprocessing can increase the accuracy obtained in reverberant conditions. The improvement in performance obtained by applying SRS is particularly significant under conditions with larger reverberation times.

Index Terms: Acoustic event detection, spectral reverberation suppression, LRSV estimation

1. Introduction

Acoustic event detection (AED) is increasingly present in applications such as health monitoring, home surveillance or security. Such applications require AED systems to be highly robust, i.e., to yield a high recognition accuracy even in the presence of acoustic disturbances. These disturbances consist of environmental noise and, in an enclosed space, of reverberation characterized by the room impulse response (RIR). Though both noise and reverberation are known to have a detrimental effect on the performance of AED systems, the work on single-channel preprocessing for AED has been focused mainly on the reduction of environmental noise [1], often using approaches initially developed in the context of speech enhancement**.** Using and adapting speech enhancement algorithms to improve the performance of a recognizer is a common approach in the field of automatic speech recognition (ASR) and we will use this approach in the context of AED. In single-channel scenarios, such algorithms often rely on spectral suppression to reduce both noise and reverberation [2]. Spectral suppression consists in applying a real valued spectral gain to the short-time Fourier transform (STFT) of the microphone signal. The computation of this spectral gain requires an estimate of the spectral variance of the interference to be suppressed. The spectral suppression scheme used in this thesis applies a spectral gain [3] computed from an estimate of the late reverberant spectral variance (LRSV) similarly as studied in the context of ASR [4]. We consider estimators of the LRSV [5,6] based on statistical models of the RIR [7, 8]. We refer to the application of a spectral gain computed from an estimate of the LRSV as spectral reverberation suppression (SRS). This paper investigates the influence of reverberation on the performance of an AED system and the potential benefit of applying SRS as preprocessing for such a system.

The remainder of this paper is structured as follows: In Section 2, the considered signal model and the spectral reverberation suppression algorithms are described. Section 3 presents the corpus used for our experiments. In Section 4 the experiments and their results are discussed. Finally conclusions are drawn in Section 5.

2. Reverberation Suppression

The microphone signal $y[n]$, at time sample n, consists of the anechoic event signal $s[n]$ emitted at the position of the source and corrupted by reverberation, characterized by the RIR $h(t)$. Through this paper, we consider that no noise source is present, and thus express the microphone signal as

$$y[n] = s[n] * h[n]. \quad (1)$$

The RIR can be considered as splitted between its early and late part, and eq. (1) can be expressed as

$$y[n] = s[n] * h_e[n] + s[n] * h_l[n] = e[n] * l[n], \quad (2)$$

with $e[n]$ and $l[n]$ denoting the early and late reverberant source components, respectively. In the STFT domain with the time-frame index l and the discrete frequency index k, Eq. (2) can be expressed as

$$Y[k,l] = E[k,l] + L[k,l], \quad (3)$$

with $Y[k,l]$, $E[k,l]$ and $L[k,l]$ denoting the STFTs of $y[n]$, $e[n]$ and $l[n]$, respectively. In the remainder of this paper, the symbol ˆ will be used to denote estimated quantities.

2.1 Spectral Gain

The application of SRS consists in applying a spectral gain $G[k,l]$ to the STFT of the input signal in order to obtain an estimate $\hat{E}[k,l]$ of $E[k,l]$ such as

$$\hat{E}[k,l] = G[k,l]\, Y[k,l]. \quad (4)$$

Assuming that $E[k,l]$ and $L[k,l]$ are uncorrelated, the spectral variance $\sigma_y^2[k,l]$ of the microphone signal can be expressed as

$$\sigma_y^2[k,l] = \sigma_e^2[k,l] + \sigma_l^2[k,l], \quad (5)$$

with $\sigma_e^2[k,l]$ and $\sigma_l^2[k,l]$ denoting the early and late reverberant spectral variances, respectively. The spectral gain of Eq. (4) can be designed as a Wiener gain, i.e.

$$G[k,l] = \frac{\sigma_e^2[k,l]}{\sigma_e^2[k,l]+\sigma_l^2[k,l]}. \quad (6)$$

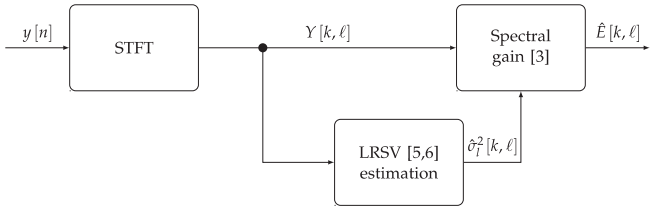

Figure 1: Overview of the application of SRS

The computation of the spectral gain $G[k,l]$ requires an estimate $\hat{\sigma}_l^2[k,l]$ of the LRSV. Thus, SRS (Fig. 1) processed based on estimating the LRSV in order to compute the spectral gain that is applied to the STFT of the input signal.

Several methods exist to compute the spectral gain. One of these is the minimum mean-square error log-spectral amplitude estimator (MMSE-LSA), which is used in the remainder of this paper. The estimate of $\sigma_y^2[k,l]$ can be obtained using recursive temporal smoothing, i.e.

$$\hat{\sigma}_y^2[k,l] = \beta\hat{\sigma}_y^2[k,l-1] + (1-\beta)|Y[k,l]|^2, \quad (7)$$

with β denoting a smoothing parameter.

2.1 LRSV estimation from a statistical model

It has been proposed in [7] to model the RIR base on a non-stationary stochastic process, i.e.

$$h(t) = \begin{cases} b[t]\exp(-\bar{\delta}t), \text{for } t > 0 \\ 0, \text{ otherwise,} \end{cases} \quad (8)$$

with $b[t]$ denoting a white zero-mean Gaussian noise signal and with $\bar{\delta}$ denoting a damping constant that is related to the T_{60} by

$$\bar{\delta} = \frac{3\ln 10}{T_{60} f_s}. \quad (9)$$

The model from Eq. (8) is valid only for DRR ≤ 0, i.e. when the distance between the source and the microphone is larger than the critical distance. This model has been generalized in [8] to take into account the cases in which DRR > 0. In this generalized model, the RIR model is split into two segments leading to

$$\sigma_y^2[k,l] = \sigma_d^2[k,l] + \sigma_r^2[k,l], \quad (10)$$

with $\sigma_d^2[k,l]$ denoting the spectral variance of the speech convolved with the direct path and $\sigma_r^2[k,l]$ the estimate of the spectral variance of all reverberations. The estimate of the reverberant spectral variance can now be expressed as a function of the estimated microphone signal, of a ratio of reverberant to direct path relation $[k]$,

$$\alpha[k] = \frac{\beta_r[k]}{\beta_d[k]}, \quad (11)$$

and the damping constant $\Delta[k]$ as,

$$\Delta[k] = \frac{3\ln 10}{T_{60}[k] f_s} \quad . \quad (12)$$

Hence, the LRSV can be estimated by

$$\hat{\sigma}_l^2[k,l] = \exp(-2\Delta[k]\ R\ (N_e-1))\ \hat{\sigma}_r^2[k,l-N_e+1], \quad (13)$$

which will be referred to as the Habets estimator. Additionally, for $\alpha[k] = 1$, the LRSV is given by

$$\hat{\sigma}_l^2[k,l] = \exp(-2\Delta[k]\ R\ (N_e-1))\ \hat{\sigma}_r^2[k,l], \quad (14)$$

which will be referred to as the Lebart estimator. The computation of the damping constant $\Delta[k]$ requires an estimate of T_{60} and the ratio $\alpha[k]$ that is related to the DRR by

$$DRR[k] = 10\ log_{10}\frac{1-\exp(-2\Delta[k]R)}{\exp(-2\Delta[k]R)}\frac{1}{\alpha[k]}. \quad (15)$$

In this work, these values were measured from the RIR.

3. Dataset

In order to conduct the experiments, a database of both anechoic and reverberant signals was recorded. To the best of our knowledge, no public database of anechoic recordings designed for the evaluation of AED exists. Therefore, acoustic events were recorded in an anechoic chamber. We were interested in events of short duration, possibly applicable to human-machine interaction (i.e. can be used to control devices), and ease of recordings. Based on these criteria, we recorded five classes of events: "clap", "clear throat", "cough", "tongue flipping" and "snap". The database of anechoic recordings was split to create a training set and a testing set. In order to create reverberant signals, the anechoic signals were convolved with recorded RIRs, from public RIRs databases. These databases were chosen to cover a wide range of RIRs parameters, i.e. in our case, wide ranges of T_{60} and DRRs.

3.1 Training Set

For the training set, a database of RIRs from three different rooms at two different distances (per room) has been used [9]. In order to obtain reverberant signals, 70% of the anechoic recordings were randomly selected and convolved with these RIRs. All RIRs correspond to a combination of a room and distance between source and microphone. In the remainder of this paper the combination of a room and a distance will be referred to as a "condition". Three different training sets were constructed from the gathered data. The first set, referred to as "anechoic training", consists of anechoic data. The second set, referred to as "1-room training", consists of reverberant data obtained by convolving the anechoic signals from the "anechoic training" set with the RIRs recorded in Room 2 at a distance of either 1 m or 2 m from the source. Therefore, the set "1-room

training" contains 2 different conditions. The third set, corresponds to multi-condition training and will be referred to as "MC training". It consists of reverberant data obtained by convolving the anechoic signals from the "anechoic training" set with the RIRs recorded in Room 1, 2 and 3 at a distance of either 1 m or 2 m from the source. Therefore, the set "MC training" contains 6 different conditions. After all the data has been convolved with the available RIRs, the set "MC training" consists of a total of 2202 utterances of acoustic events.

3.1 Testing Set

The testing set was generated using the remaining 30% of the anechoic recordings and convolving them with a second database of RIRs [10]. All testing conditions and their characteristics are summarized in Tab. 4.3, along with their corresponding T_{60} and DRR. After all the data has been convolved with the available RIRs, the testing set consists of a total of 7728 utterances of acoustic events. All signals were sampled at a sampling frequency of 16 kHz. The STFTs were computed using a Hamming window with a length of 25 ms and an overlap of 15 ms. The necessary estimation of the T_{60} and of the DRR for applying SRS, was made by measuring them from the RIRs. The T_{60} is measured by the Schroeder method [11]. The DRR has been measured using eq. (14). The AED system used 39 MFCCs (including deltas and double-deltas) per frame and three states HMMs with ten Gaussian components per state.

4. Experimental Results

4.1 Influence of reverberation on AED performance

To evaluate the influence of reverberation on the performance of an AED system and the potential benefits of training on reverberant data the AED system is applied to the (unprocessed) data from the testing set using statistical acoustic models constructed from the three different training sets. In addition, the scores obtained when using anechoic data for both training and testing will be presented and denoted by "optimum". The "optimum" represents the performance that can be obtained if perfect dereverberation was achieved. In this experiment, the performance of the AED system is measured in terms of accuracy, defined as

$$\text{Accuracy} = \frac{\text{Number of correctly identified events}}{\text{Number of events to be recognized}} \cdot 100\ \%.$$

The results, in terms of accuracy, are depicted in Fig. 2, as a function of either the T_{60} (top) or the DRR (bottom).

The depicted scores correspond to the accuracy obtained on the testing set, of reverberant data, when using statistical models trained on the three considered training sets. The scores labeled as "optimum" yield the highest accuracy, as expected, illustrating that the presence of reverberation is detrimental to the performance. This detrimental effect is particularly noticeable for higher values of T_{60} and lower values of DRR. This degradation in accuracy is particularly severe in the case "anechoic training", where the accuracy decreases as low as 20 %, which in our five classes recognition task corresponds to the chance level. The scores achieved using "1-room training" and "MC training" illustrate the increase in robustness that can be obtained by training on reverberant data. Both training sets yield similar improvement, with an improvement of 20 to 40 % compared to the performance obtained using "anechoic training".
The small difference in performance between "1-room training" and "MC training" could be explained by the similarity of both training sets. A slightly higher accuracy is obtained using "MC training" for conditions with higher values of T_{60}> 3 s.

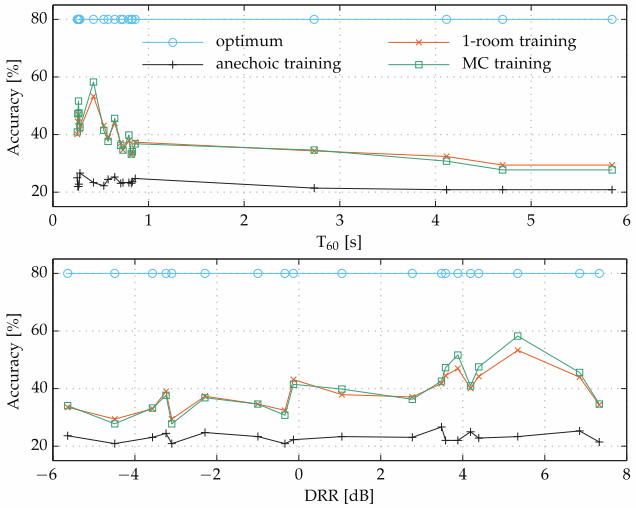

Figure 2: Accuracy as a function of the T_{60} (top) and DRR (bottom) obtained on reverberant data, using the three different considered training sets. Optimum denotes the use of anechoic data for both training and testing. An accuracy of 20 % corresponds to the chance level.

4.2 Application of SRS as preprocessing for AED

This experiment examines the benefit on the performance of an AED system when applying SRS. Here, SRS is applied using an estimate of the LRSV obtained using either the Lebart [5] estimator or the Habets estimator [6]. The required estimates of the T_{60}, for the Lebart estimator, and of both T_{60} and DRR, for the Habets estimator, have been measured from the RIRs. The same SRS scheme has always been applied to both training and testing data, except for "anechoic training" for which the anechoic data remained unprocessed. In this experiment, the performance of the AED system is measured in terms of relative improvement (Rel. Imp.), compared to the performance of the system obtained during the experiment described in Section 4.1. The relative improvement, compared to the results from Fig. 2, is depicted in Fig. 3, as a function of the T_{60}. These scores correspond to the relative improvement obtained when applying SRS using either the Lebart or the Habets estimator of the LRSV and considering the first 48 ms of the RIR that correspond to the early reflections. A limited improvement is noticable for "anechoic training" when applying SRS.
Since both LRSV estimators are based on similar models, similar results could have been expected. However, this is not the case when using the training sets "1-room training" and "MC training" where a considerable difference can be observed. The best performance, obtained using the Habets estimator, is observed in the case of "anechoic training", in which the highest relative improvement (around 35 %) is

obtained for low values of T_{60}. For higher values of T_{60} the improvement observed is negative for "1-room training" and "MC training" when using the Lebart estimator of the LRSV. The Habets estimator performs well in this case. For $T_{60} > 4$ s, a higher relative improvement is obtained. By using SRS as preprocessing, an increase in the performance of the considered AED system can be obtained, but only if using an adequate estimator for the LRSV.

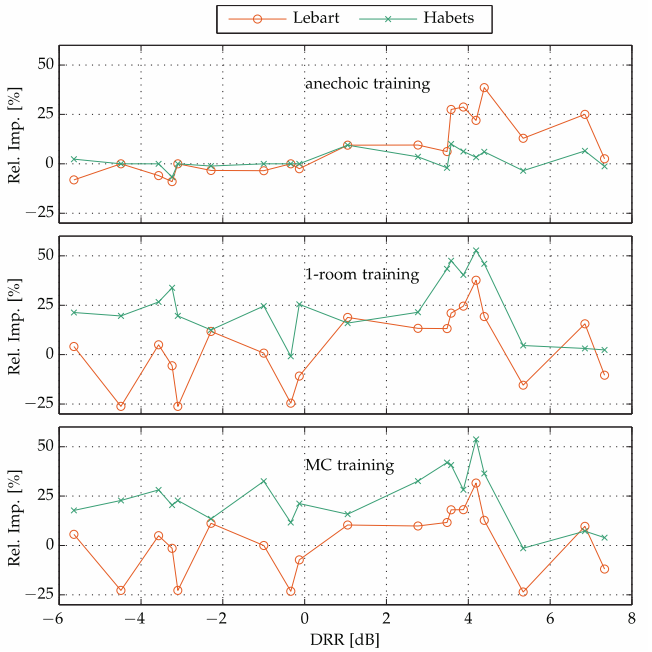

Figure 3: Relative improvement as a function of the T_{60}, for three training sets, between accuracies obtained using unprocessed data (cf. Fig. 2) and data processed using SRS using either the Lebart of the Habets estimator of the LRSV. In the case of the anechoic training set, only the testing data has been processed.

5. Summary and Conclusion

The performance of the considered AED system has been evaluated for three different training sets, constructed using a database of anechoic acoustic events and publicly available databases of RIRs. The results demonstrated that reverberation is detrimental to the performance of an AED system, even in conditions with short reverberation times. It appeared as well that this deterioration in performance gets higher for higher values of T_{60} and lower values of DRR. Indeed, when training on anechoic data the "optimum" performance (testing on anechoic data) yields an accuracy of 80 % while testing on reverberant data yields an accuracy close to 20 %, i.e. the chance level in the considered five classes classification task. Training on reverberant data, i.e. "1-room training" and "MC training" through this thesis, improved the robustness to reverberation by up to 40 % (relative improvement) compared to training on anechoic data. An SRS scheme has been applied as preprocessing in order to improve the accuracy in reverberant conditions. This SRS scheme consists in the application of a spectral is based on a gain computed from an estimate of the LRSV. Two estimators of the LRSV have been considered, referred to as either the Lebart or the Habets estimator. It appeared that when using the Lebart estimator, SRS could actually decrease the performance of the AED system. However, when using the Habets estimator, the performance is increased in all conditions, with a relative improvement up to 50 % compared to the accuracy obtained on unprocessed reverberant data. However, the obtained results suggest that training on reverberant data is still beneficial for robustness against reverberation, though the application of SRS can lead to an additional improvement of the accuracy.

Acknowledgements

This work was partially funded by the European Commission Grant no. 318381 EAR-IT Experimenting Acoustics in Real environments using Innovative Test-beds, S4CoB – Sound for Energy Control of Buildings under grant no. 284628 and EcoShopping – Energy efficient and Cost competitive retrofitting solutions for shopping buildings, grant no. 609180.

References

[1] J. Schröder, N. Moritz, M.R. Schädler, B. Cauchi, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze. On the use of spectro-temporal features for the IEEE AASP challenge “Detection and Classification of Acoustic Scenes and Events”. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, N.Y., U.S.A., October 2013.

[2] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Juki´c, T. Gerkmann,S. Doclo, and S. Goetze. Joint dereverberation and noise reduction using beamforming and a single-channel speech-enhancement scheme.In Proc. REVERB challenge workshop, Florence, Italy, May 2014

[3] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process., 33(2):443–445, April 1985.

[4] R. Maas. One the application of reverberation suppression to robust speech recognition. IEEE Trans. Speech Audio Process., pages 297–300, 2012.

[5] K. Lebart, J. M. Boucher, and P. N. Denbigh. A new method based on spectral subtraction for speech de-reverberation. Acta Acoustica, 87:359–366, 2001.

[6] E. A. P. Habets, S. Gannot, and I. Cohen. Late reverberant spectral variance estimation based on a statistical model. IEEE Signal Process. Lett., 16(9):770–774, September 2009.

[7] J. D. Polack. Playing billiards in the concert hall: the mathematical foundations of geometrical room acoustics. Applied Acoustics, 38(2):235–244, 1993.

[8] E.A.P. Habets. Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement. Ph.D. Thesis, june 2007.

[9] E. Hadad, F. Heese, P. Vary, and S. Garnnot. Multichannel audio database in varoius acoustic environment. Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC), 2014.

[10] M. Jeub, M. Schäfer, and P. Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. pages 1–4, Sydney, Australia, July 2009.

[11] M. R. Schroeder. New method of measuring reverberation time. J. Acoust. Soc. Am., 37:409–412, 1965.